Received July 29, 2021, accepted August 27, 2021, date of publication September 6, 2021, date of current version September 29, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3110604

A Knowledge-Based Clinical Decision Support System Utilizing an Intelligent Ensemble Voting Scheme for Improved Cardiovascular Disease Prediction

SABA BASHIR[1], ABDULWAHAB ALI ALMAZROI[2], SUFYAN ASHFAQ[1], ABDULALEEM ALI ALMAZROI[3], AND FARHAN HASSAN KHAN[4]

[1]Department of Computer Science, Federal Urdu University of Arts, Science and Technology, Islamabad 46000, Pakistan
[2]Department of Information Technology, College of Computing and Information Technology at Khulais, University of Jeddah, Jeddah 21959, Saudi Arabia
[3]Department of Computer Science, Applied College, Northern Border University, Arar 91431, Saudi Arabia
[4]Knowledge and Data Science Research Center (KDRC), Department of Computer and Software Engineering, College of E&ME, National University of Sciences and Technology (NUST), Islamabad 44000, Pakistan

Corresponding author: Saba Bashir (saba.bashir3000@gmail.com)

ABSTRACT A massive amount of medical data is available in healthcare industry, which can be utilized to extract useful knowledge. A Clinical Decision Support System (CDSS) is used to improve patient's safety by minimizing medical errors. Heart disease is one of the major chronic maladies even in todays' world. Many researchers have employed different data mining techniques to predict heart disease. The objective of proposed framework is to improve the accuracy of heart disease prediction. In this paper, an ensemble based voting scheme is proposed to efficiently predict heart disease. Four benchmark heart disease datasets from UCI repository have been utilized for experimentation and evaluation. The performance of the proposed ensemble is compared with individual classifiers as well as with five different ensemble schemes using various parameters in order to show the effectiveness of the proposed ensemble scheme. The evaluation of results shows that the proposed ensemble scheme has better average accuracy (83%) as compared to other ensemble schemes as well as individual classifiers.

INDEX TERMS Data mining, clinical decision support system, ensemble scheme, machine learning classifiers.

I. INTRODUCTION

Medical organizations and hospitals are generating large amount of medical data on daily basis; however this data cannot be used intelligently until useful knowledge is extracted. Data Mining can be used to discover specific hidden information from large raw datasets [1]. Medical data mining is one of the most active areas of research to find interesting patterns and meaningful information from medical data sets. Intelligent data mining techniques can be applied on these data sets to convert them into useful information [2].

Data mining has been used effectively in different fields such as marketing, businesses and banking etc. Its benefits can most importantly be seen in health care since it has been useful in predicting diseases such as heart attack [3], breast cancer [4], diabetes [5] and hepatitis [6] with high accuracy. Nowadays heart diseases or cardiovascular diseases (CVDs) have become a very hot issue globally. Heart disease is one of the major causes of deaths worldwide which is increasing rapidly with the passage of time [2], [7]. Detection of heart attack at an earlier stage can reduce the chances of death and other severe consequences [8]. According to a survey conducted by World Health Organization (WHO), cardiovascular diseases have become a great cause of death and around 17.7 million people died in 2015 which makes CVDs to be the cause of 31% of the total deaths. It was estimated that 7.4 million of people died due to heart diseases and 6.7 million deaths were due to stroke [9]. According to world heath rankings, Pakistan is ranked 63rd in the world where the rate of deaths caused by cardio vascular diseases is 110.65 per 100,000 [10].

The associate editor coordinating the review of this manuscript and approving it for publication was Mu-Yen Chen.

TABLE 1. A summary of state of the art approaches for cardiovascular diseases prediction.

Author/Reference	Year	Techniques	Specificity %	Sensitivity %	Accuracy %
Das et al. [3]	2009	NN Ensemble	95.91	80.95	89.01
Srinivas et al. [18]	2010	NB	--	--	83.96
Ghumbre et al. [21]	2011	SVM RBF	88.50 82.10	84.06 82.40	85.05 82.24
Chen et al. [16]	2011	ANN	70	85	80
Peter & Somasundurm. [15]	2012	NB KNN DT NN	--	--	82 74 69 78
Chitra & Sinivasagam. [19]	2013	CNN SVM	87 77.5	83 85.5	85 82
Bashir et al. [24]	2014	Ensemble Vote	92.86	73.68	81.82
Ismaeel et al. [26]	2015	Extreme Machine Learning (EML)	--	--	80
Bashir et al. [11]	2015	Ensemble Bagging	82.87	61.23	81.87
Verma et al. [25]	2016	FURI MLP MLR C4.5	--	--	87.05 90.28 91.36 85.6
Khateeb et al. [20]	2017	K Nearest Neighbor	--	--	80

In the prediction of CVDs, accuracy plays a vital role as an individual's life is at stake. Medical errors are responsible for deaths throughout the world, even a single medical error can lead to sudden death. These medical errors can effectively be reduced by more accurate data mining techniques for disease prediction. Several data mining techniques have been introduced by researchers to improve the accuracy in medical health community.

Clinical decision support system (CDSS) plays an important role for automated diagnosis of heart disease. Such systems are developed using data mining techniques. CDSS provides necessary knowledge to diagnose/predict any disease with high accuracy [11], [12]. CDSS can be divided into two main categories: 1) Knowledge-based CDSS. 2) Non-knowledge based CDSS [13]. A knowledge base is used Knowledge-based CDSS where clinical rules are used intelligently to form the knowledge. These rules can be ''if then statements'' or inference rules whereas, Non-knowledge based CDSS displays results of patients' clinical data in a simplified manner. The proposed research focuses on knowledge-based CDSS where a heart disease diagnosis framework is proposed that results in high accuracy of heart disease prediction.

Following are some main contributions of the proposed research:

- A novel combination of machine learning classifiers is proposed in ensemble voting schemes to predict heart disease
- Empirical evaluation of individual classifiers for heart disease prediction and then performance-based selection of individual classifiers for ensemble voting schemes
- Proposed different ensembles with novel combinations of selected classifiers and performed their evaluation to show high performance results
- Specifically, 4 heart disease datasets are utilized for performance evaluation of 6 single classifiers and 5 ensemble voting schemes
- Proposed ensemble approach has achieved better results as compared to individual and other ensemble schemes

Rest of the paper is arranged as follows: Section 2 discusses state of the art literature review related to heart disease diagnosis. Section 3 presents proposed approach, description of datasets and working of proposed ensemble. Evaluations and measures are presented in Section 4 whereas Section 5 elaborates the results and discussion. Finally, conclusions and future work are given in Section 6.

II. LITERATURE REVIEW

In the past few decades, different data mining techniques have been used for designing clinical decision support systems. An overview of state of the art literature review for heart disease diagnosis is presented in Table 1 and described in detail in this section. In [14] Pattekari and Parveen proposed a system that used Naïve Bayes to predict heart diseases. The proposed research worked only for the categorical data. Using other data mining techniques, this technique can be improved and better outcome may be achieved by examining other data types. Peter and Somasunduram [15] put forward a method using data mining and pattern recognition techniques for heart disease diagnosis. It is analyzed that performance of Naïve Bayes is comparatively improved and generates better results. However, this technique restricts the use to only numerical attributes set and encourages the ASCII file format only. Chen *et al.* proposed a framework in [16] that employed Learning Vector Quantization (LVQ) algorithm for heart disease prediction. This algorithm used Receiver Operating Characteristics (ROC) curve to show the results

and 80% accuracy was achieved. It can be enhanced by using text mining alongside data mining techniques, as text mining has the ability to mine unstructured data that can be used in the heart disease prediction.

Association rule mining and Genetic Algorithm (GA) based approach was proposed by Jabbar *et al.* [17]. These techniques achieved high values of accuracy and interestingness measures for the prediction of heart disease. However, this framework made use of the whole feature space for training where improvements can be applied by using feature selection technique to reduce the feature space. In [18] Srinivas *et al.* proposed a model which can give response to complex queries. The proposed framework utilized different classifiers in order to perform heart disease prediction. The approach is efficient as compared to others because it uses training data based on only 15 attributes. High efficiency and reduced training time is achieved by the proposed framework. Chitra and Sinivasagam [19] also used machine learning classifiers for the prediction of heart disease at an early stage. The high specificity indicated that the correct patient is predicted healthy whereas the high sensitivity shows that the patient is appropriately predicted that he has high chances of having the disease. Khateeb *et al.* [20] used kNN classifier and achieved an accuracy of 80%. In [21] Ghumbre *et al.* presented a framework for diagnosis of heart disease. Support Vector Machine (SVM) and Radial based Functions (RBF) network are used for model construction. Their analysis demonstrated that the results obtained from SVM are as good as RBF.

A lot of work has been done in literature on single classifiers. However, to overcome the limitations of single classifiers, ensemble schemes have been introduced. Ensemble approaches consist of multiple classifiers and may be used to improve the accuracy. These schemes have been used by different authors to reach the desired level of accuracy. In [3] Das *et al.* proposed an ensemble based approach for heart disease diagnosis. This approach used a combination of Neural Networks, trained on same type of data. Only one dataset has been used in this technique by the authors. More datasets maybe utilized for the verification of results. In [22] Helmy *et al.* proposed an ensemble based approach using SVM, ANN and ANFIS for more accurate prediction of heart disease. Bagging algorithm has been used to train the individual classifiers. The results show that heterogonous classifiers performed better as compared to individual classifiers. However, this approach made use of only two data sets. Maroco *et al.* [23] presented an approach to improve performance by applying neuro-psychological testing. Bashir *et al.* [24] also proposed an ensemble vote scheme using NB, DT and SVM for prediction of heart disease and achieved 81.2% accuracy. In [25] Verma *et al.* put forward a hybrid approach using Fuzzy Unordered Rule Induction Algorithm (FURI), Multilayer Perceptron (MLP) and Multinomial Logistic Regression (MLR). This approach was evaluated only on one data set. In [26] Ismaeel *et al.* put forward an Extreme Machine Learning (EML) algorithm for the diagnosis of heart disease. The proposed method achieved an accuracy of 80%. In [11] Bashir *et al.* proposed an ensemble using Bootstrap Aggregation (Bagging). The proposed ensemble achieved an average accuracy of 81.87%. However the authors have not measured the performance in terms of correlation, classification error, absolute error, relative error and kappa statistics.

[27] proposed a method on medical datasets for disease diagnosis. The Bayesian network based method is used to deal with overconfidence. In [28] decision tree learning method is adopted for clinical data. The proposed method effectively handled noisy data and generated high performance results. [29] also proposed a method which works for medical data and perform structural and textual information fusion. Heart disease identification method is also proposed in [30] using novel feature selection and classification algorithms. [31] also uses machine learning classification algorithms for medical datasets. Feature extraction is performed using convolution neural networks and then data is classified into diseased and healthy classes.

An overview of state of the art literature for heart disease prediction indicates that this research area is prime focus of the research community. Although, huge research has been conducted for the prediction of heart diseases, but there is still space for improvement as some approaches work only on one data type while others are evaluated only on a single dataset. There are several others which do not achieve the acceptable levels of accuracy. Hence, this research gap presents an opportunity to propose an approach that achieves high performance results for heart disease prediction on heterogeneous attribute types and the verification of the technique on multiple benchmark datasets.

III. PROPOSED APPROACH

The proposed approach is described in detail in this section. Each component is discussed in detail.

A. DATA ACQUISITION AND PREPROCESSING

Benchmark dataset are obtained from heart disease repository which is freely available online. The proposed technique uses 4 benchmark heart disease datasets taken from UCI data repository [32] namely Cleveland, Hungarian, Long Beach and Switzerland dataset. Each dataset contains different set of attributes. Classifiers are trained on these datasets individually and ultimately trained classifiers are used to determine the presence/absence of disease. Four different datasets are used to show the diversity of proposed model.

After data acquisition, data preprocessing is applied to refine the datasets i.e missing values imputation, outlier detection and removal. Feature selection is also applied to identify the most relevant attributes for disease classification.

B. MAJORITY VOTING BASED ENSEMBLE SCHEMES

The proposed approach is comprised of novel ensemble scheme. Five novel ensembles have been introduced by using the base classifiers. The base classifiers are selected showing high performance results based on literature. Multiple

ensembles are introduced as they show high performance on different heart disease datasets and then ensemble with highest average accuracy, sensitivity, specificity and F-measure is selected for the final classification.

The proposed approach uses five different majority voting based ensemble schemes and their performances are analyzed. The proposed technique has the following important steps: First step is to generate the classification decisions of independent classifiers for each heart disease dataset. Second step involves computation of the average results of individual classifiers and select the top 3 on the basis of average accuracy. In third step, top-3 individual classifiers are combined in ensemble voting schemes and their results are evaluated. The performance of the selected ensemble schemes for all the heart disease dataset is noted. Finally, in the fourth step, the average results of ensemble vote schemes, across all dataset, are computed and compared. The novelty has been introduced here as the combination and flow/working of proposed methodology is not existed before. Figure 1 represents detailed architecture of proposed framework. The proposed framework utilized different machine learning classifiers for the classification task. The description of the each classifier is given below:

1) NAÏVE BAYES

Naïve Bayes classifier considers each attribute independently to determine the presence of disease. It uses only a small dataset for training purpose. Furthermore, it only requires the class attribute, as other attributes do not depend upon each other [12] . The formula used for NB classifier is given below:

$$P(C_K|X) = \frac{P(C_K) * P(X|C_K)}{P(X)} \tag{1}$$

where x is a predictor, C_K is probability of a class and $P(C_k|X)$ is the probability of C_K class given attribute X.

2) SUPPORT VECTOR MACHINE

The basic property of SVM is that it performs binary classification. For each input set a prediction model is built and output is produced in the form of two classes. An SVM model is therefore a presentation of an example in which points in space are mapped in a way so that the points in other examples of different categories could be mapped as wide as possible. After that, new examples are mapped in the same space and the side they belong to is estimated on the basis of category [2].

SVM uses the following classification rule for solving the problem:

$$\text{sgn}(f(X, W, b)) \tag{2}$$

$$f(X, W, b) = \langle w.x \rangle + b \tag{3}$$

where x is the example that needs to be classified, whereas f(X, W) represents a complex problem for maximum margin hyper plane. We have classified the data sets into two classes (Healthy, Sick) by using heart disease dataset attributes.

3) DECISION TREE

Decision Tree (DT) is used for data cleaning and pattern reorganization. A dataset may have a large number of attributes that are less important for the research process, so dataset attributes can be reduced by using Gini Index. DT classifier uses graph like structure and does not require knowledge of domain. Conditional probability of each node is calculated and is used to select the best alternative. Decision trees are widely used in clinical decision support systems and for disease diagnosis. Attributes are selected with the lower Gini Index and then rules are generated from the selected attributes [11]. Gini Index is calculated by using the following formula:

$$Gini(t) = 1 - \sum_{i=1}^{c-1} [p(i, t)] \tag{4}$$

where total no. of classes are denoted by c-1 and p(i, t) denotes the i class probability in t class.

4) NEURAL NETWORK

Neural Network is an inspired network form human brain. It has unbelievable processing ability owing to the massive network of interconnected neurons. It has three layers. Input layer provides an interface, hidden layer is used for computation and output layer stores the output [3]. Model training is performed by Back propagation algorithm. It initializes the weights by using small random number after which it trains the input data. In the third step it computes the output for each unit by using sigmoid function equation, given below:

$$o = \sigma(\vec{w}, \vec{x})\sigma(y) = \frac{1}{1 + e^{-y}} \tag{5}$$

where $\vec{w}$ denotes unit values vector and $\vec{x}$ denotes weights values vector.

The error calculation steps come after that. The error rate ∂ is transmitted to all neurons and is calculated for each network output. For each Output unit k, error term ∂_k is calculated by using equation:

$$t_k \leftarrow o_k(1 - o_k)(t - t_k) \tag{6}$$

where o_k represents output for unit k.

For each hidden unit h, error term ∂_h is calculated by using equation.

$$t_k = o_h(1 - o_h) \sum_{k \in outputs} w_{kh}\partial_k \tag{7}$$

where w_{kh} represents the weight of network unit k from hidden unit h to k.

Upgrade each network weights as given below:

$$w_{ji} \leftarrow w_{ji} + \Delta w_{ji} \tag{8}$$

where $\Delta w_{ji} = \eta \partial_j x_{ij}$, η denotes the learning rate and x_{ij} represents the represents the inputs from unit I to unit j.

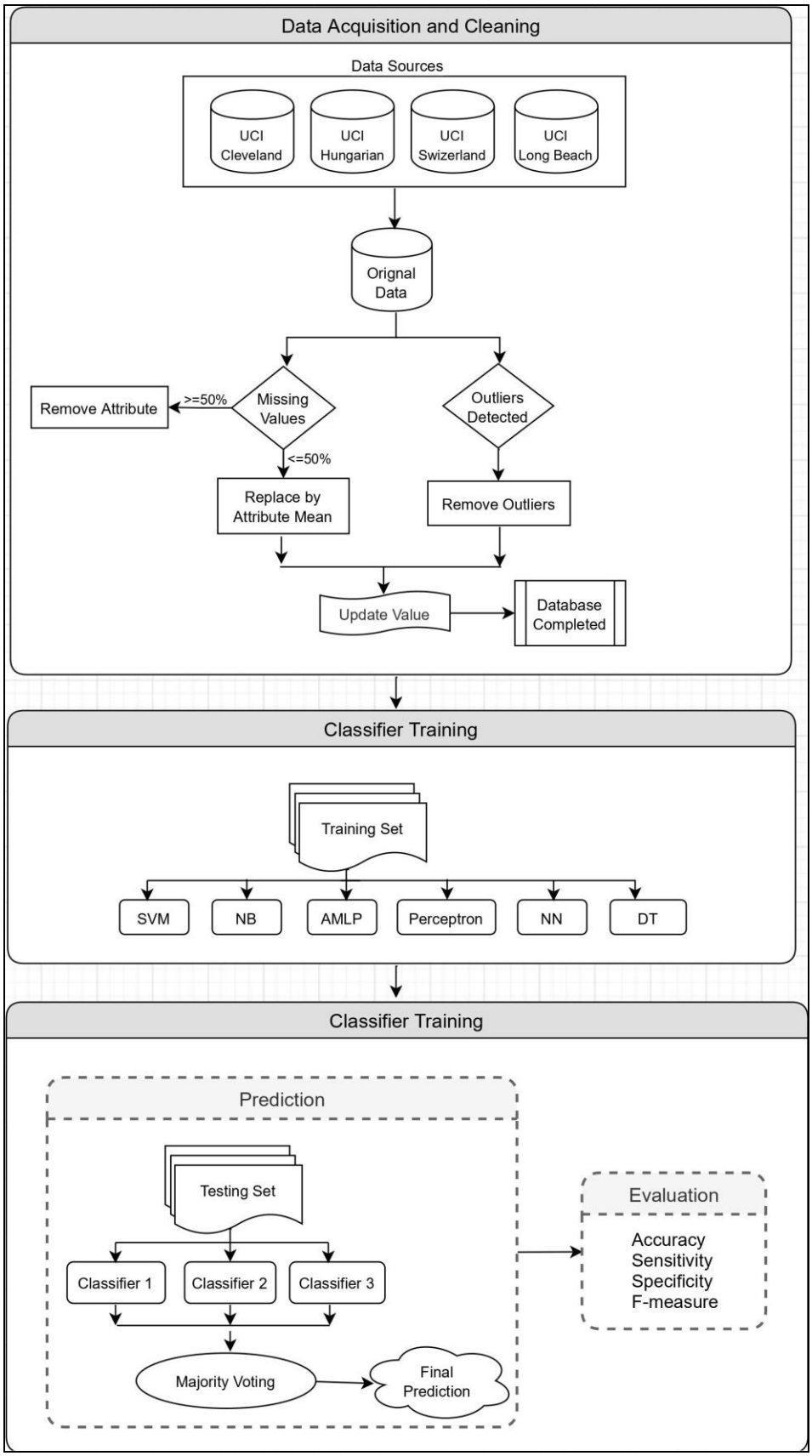


FIGURE 1. Detailed architecture and flow of proposed framework.

5) MULTILAYER PERCEPTRON

The mapping from input nodes to output nodes is performed by Multilayer Perceptron (MLP). It is a feed forward neural network. Auto MLP has a directed graph that contains number of nodes arranged in layers pattern and each node is connected with one another. Auto MLP structure has more than one hidden layers for solving complex problems which could not be solved by a single hidden layer [25]. Auto MLP makes use of Back Propagation Algorithm for the training of network that has already been discussed in previous section.

6) SINGLE LAYER PERCEPTRON

Perceptron has a single layer of output nodes so it can be considered as a simple kind of Feed Forward Neural Network which can only classify linear separable cases in the form of binary target (0, 1). It doesn't have a prior knowledge so the initial weights are assigned to the input layer randomly. Single Layer Perceptron sums up all the weights. If the sum of weights is greater than threshold (Pre-determined Value) then the single layer is considered to be active. The input values of the weights are provided to the Perceptron, if the predicted output is same as the expected output then performance is considered to be satisfactory. Otherwise, weights need to be assigned again to minimize the error [34] . Its algorithm concept is described as below:

$$w_1x_1 + w_2x_2 + \ldots\ldots\ldots w_nx_n \rangle threshold \rightarrow Active \quad (9)$$

$$w_1x_1 + w_2x_2 + \ldots\ldots w_nx_n \leq threshold \rightarrow NotActive \quad (10)$$

C. WORKING OF PROPOSED FRAMEWORK

Working of proposed framework is divided into two major phases. The first phase consists of data acquisition and pre-processing. Data acquisition is performed by obtaining the heart disease data from available resources. In order to support comparison of this research, 4 benchmark heart disease datasets from UCI repository have been utilized. Data pre-processing is applied to the acquired data which involves outlier detection, outlier removal and replacement of missing values. Outliers are the values that fall below or above a specific range and they have been filtered out in this research. Missing values in dataset may have a serious effect on the heart disease prediction so they should be handled carefully. In the proposed system, missing values have been replaced by the mean of the specific attribute. If a certain attribute has more than 50% of the values that are missing then that attribute has been discarded.

The second phase of the proposed approach applies majority voting based ensemble using different individual classifiers. Novel combination of classifiers along with novel method to compute final ensemble is introduced. In majority voting each classifier within the ensemble produces a class and the class having most votes is the one that is suggested by the ensemble. This phase starts off by computing accuracy of individual classifiers on the acquired and pre-processed data. The results are computed and top-3 classifiers are identified for each dataset based on accuracy. As there are four datasets, this may result upto four different combinations of individual classifiers. These four different combinations of individual classifiers are combined using majority voting based ensemble. Each ensemble vote includes three heterogeneous individual classifiers. For Cleveland dataset, the proposed approach results in a combination of SVM, NB and AutoMLP called Ensemble 1. Similarly for Hungarian dataset SVM, NB and NN are combined and called Ensemble 2. For Switzerland dataset Ensemble 3 is composed of SVM, Perceptron and NN. Lastly, for Long Beach data set DT, NB and AutoMLP are combined and called Ensemble 4. The fifth majority vote based ensemble scheme, called Ensemble 5, is constructed by using the average results of four heart disease dataset in terms of accuracy and selecting the top 3 individual classifiers i.e. SVM, NN and AutoMLP. These five ensemble vote schemes are then trained & tested on four heart disease data set and the results are computed.

TABLE 2. Cleveland dataset attributes name and description.

Attribute Names	Description
Age	Age in years
Sex	Sex
CP	Chest pain type
Trestbps	Resting blood pressure
Chol	Serum cholesterol in mg/dl
Fbs	Fasting blood sugar > 120 mg/dl
Restecg	Resting electrocardiographic results
Thalach	Maximum heart rate achieved
Exang	Exercise induced angina
Oldpeak	ST depression induced by exercise relative to rest
Slope	Slope of the peak exercise ST segment
Ca	Number of major vessels colored by flourosopy
Thal	3=Normal; 6= fixed defect; 7= reversible defect

TABLE 3. Confusion matrix.

Predicted Class	Actual Class	
	X	Y
X	True Positives	False Negatives
Y	False Positives	True Negatives

IV. EXPERIMENTAL EVALUATION AND MEASURES

A comprehensive elaboration of results in given in this section. Standard 10-fold cross validation has been used to

TABLE 4. Performance comparison of individual classifiers for Cleveland dataset.

Sr. no	Techniques	Accuracy	Sensitivity	Specificity	F-Measure	Classification Error	Kappa	Absolute Error	Relative Error	Correlation
1	Neural Network	78.53	78.35	78.49	78.41	21.47	0.565	0.218	21.83	0.578
2	Decision tree	74.10	73.97	73.58	73.77	25.90	0.475	0.289	28.86	0.486
3	Naïve Bayes	83.68	83.68	83.26	83.46	16.32	0.669	0.185	18.49	0.678
4	SVM	79.90	82.36	78.57	80.42	20.10	0.584	0.274	27.36	0.616
5	Perceptron	55.60	61.94	58.12	59.96	44.40	0.162	0.443	44.32	0.230
6	AutoMLP	81.61	81.42	81.49	81.45	18.39	0.630	0.214	21.38	0.645

TABLE 5. Performance comparison of individual classifiers for Hungarian dataset.

Sr. no	Techniques	Accuracy	Sensitivity	Specificity	F-measure	Classification error	Kappa	Absolute Error	Relative Error	Correlation
1	Neural Network	74.94	72.65	72.94	72.79	25.06	0.031	0.261	26.07	0.031
2	Decision tree	64.42	61.77	62.28	62.02	35.58	-0.066	0.400	39.97	0.000
3	Naïve Bayes	81.74	79.87	81.07	80.47	18.26	0.018	0.208	20.75	0.018
4	SVM	82.08	82.86	77.01	79.82	17.92	0.579	0.250	25.04	0.029
5	Perceptron	34.48	17.37	49.00	25.64	65.52	-0.014	0.655	65.51	0.000
6	AutoMLP	72.08	69.73	70.31	70.01	27.92	0.004	0.303	39.29	0.004

TABLE 6. Performance comparison of individual classifiers for Switzerland dataset.

Sr. no	Techniques	Accuracy	Sensitivity	Specificity	F-measure	Classification error	kappa	Absolute Error	Relative Error	Correlation
1	Neural Network	91.21	60.59	63.86	62.18	8.79	0.240	0.094	9.39	0.152
2	Decision tree	90.30	47.22	47.66	47.44	9.70	-0.051	0.111	11.08	0.000
3	Naïve Bayes	64.47	54.95	73.44	62.86	35.33	0.104	0.357	35.67	0.158
4	SVM	94.85	47.34	50.00	54.18	5.15	0.000	0.299	29.91	0.000
5	Perceptron	93.03	47.29	49.06	48.11	6.97	-0.027	0.074	7.35	0.000
6	AutoMLP	89.39	53.87	55.06	54.46	10.61	0.087	0.119	11.86	0.052

TABLE 7. Performance comparison of individual classifiers for Long Beach dataset.

Srno	Techniques	Accuracy	Sensitivity	Specificity	F-measure	Classification error	Kappa	Absolute Error	Relative Error	Correlation
1	Neural Network	74.94	72.65	72.94	72.79	25.06	0.031	0.261	26.07	0.031
2	Decision tree	64.42	61.77	62.28	62.02	35.58	-0.066	0.400	39.97	0.000
3	Naïve Bayes	81.74	79.87	81.07	80.47	18.26	0.018	0.208	20.75	0.018
4	SVM	82.08	82.86	77.01	79.82	17.92	0.579	0.250	25.04	0.029
5	Perceptron	34.48	17.37	49.00	25.64	65.52	-0.014	0.655	65.51	0.000
6	AutoMLP	72.08	69.73	70.31	70.01	27.92	0.004	0.303	39.29	0.004

divide the data into training and testing sets. The missing values in datasets have been replaced by the mean of total values for the specific feature by using 'Replace Missing Values' operator and the outliers have been detected and filtered out by the 'filter' operator. The description of each dataset is given as follows.

TABLE 8. Average accuracy, sensitivity, specificity and f-measure for four datasets.

Sr.	Techniques	Accuracy (%)	Sensitivity (%)	Specificity (%)	F-Measure (%)
1	Neural Network	77.35	65.66	66.55	66.07
2	Decision tree	75.62	60.04	58.62	59.29
3	Naïve Bayes	76.15	71.01	74.86	72.58
4	SVM	81.97	63.86	53.53	65.01
5	Perceptron	63.54	45.01	52.22	46.69
6	AutoMLP	78.53	66.21	66.34	66.31

TABLE 9. Comparison of ensembles schemes for Cleveland dataset.

Sr	Techniques	Accuracy	Sensitivity	Specificity	F-measure	Classification error	Kappa	Absolute Error	Relative Error	Correlation
1	Ensemble 1 (SVM+NB+AutoMLP)	84.00	84.80	83.22	84.00	16.00	0.673	0.182	18.16	0.688
2	Ensemble 2 (SVM+NB+NN)	82.30	82.78	81.59	82.18	17.70	0.638	0.190	18.95	0.651
3	Ensemble 3 (SVM+Perceptron +NN)	80.94	80.74	80.74	80.74	19.06	0.614	0.283	28.31	0.629
4	Ensemble 4 (DT+NB+AutoMLP)	83.32	83.29	82.94	83.11	16.68	0.661	0.202	20.20	0.671
5	Ensemble 5 (SVM+NN+AutoMLP)	81.63	81.74	81.08	81.40	18.37	0.626	0.201	20.10	0.637

TABLE 10. Comparison of ensembles schemes for Hungarian database.

Sr. no	Techniques	Accuracy	Sensitivity	Specifi -city	Fmeasure	Classification error	Kappa	Absolute Error	Relative Error	Correl ation
1	Ensemble 1 (SVM+NB+Auto MLP)	83.49	82.30	80.83	81.56	16.51	0.018	0.204	20.41	0.019
2	Ensemble 2 (SVM+NB+NN)	82.07	80.96	78.83	79.88	17.93	0.025	0.199	19.94	0.027
3	Ensemble 3 (SVM+Perceptron+N N)	78.50	76.48	77.26	76.86	21.50	0.038	0.357	35.69	0.038
4	Ensemble 4 (DT+NB+AutoMLP)	78.15	76.07	76.07	76.07	21.85	0.011	0.273	27.25	0.011
5	Ensemble 5 (SVM+NN+AutoML P)	82.41	81.63	78.88	80.23	17.59	0.600	0.205	20.53	0.027

A. CLEVELAND DATASET

Cleveland heart disease dataset obtained from the repository include 76 attributes in total; however the proposed framework makes use of only 14 of them to get more accurate results which have already been identified by the dataset providers. The description of the dataset attributes is given in Table 2.

This dataset consists of 303 instances and has only one attribute with missing values. These values have been replaced by mean of total values for that particular attribute.

The Cleveland dataset has already assigned 4 values (0, 1, 2, 3, and 4) to the goal attribute named ''num'' where 0 means absence of disease and implies that the patient is healthy and other values (1-4) denote the presence of disease indicating the patient to be sick.

B. HUNGARIAN DATASET

Hungarian dataset includes 14 attributes and 294 instances. Many rows of this dataset contain missing values that have been replaced. This dataset has assigned 2 values (0 and 1) to the goal attribute named ''num'' where 0 means absence of disease and implies that the patient is healthy and 1 denotes the presence of disease indicating the patient to be sick.

C. SWITZERLAND DATASET

Switzerland dataset has 14 attributes and 123 instances. The Switzerland dataset has assigned 4 values (0, 1, 2, 3, and 4) to the goal attribute named ''num'' where 0 means absence of disease and implies that the patient is healthy and other values denote the presence of disease indicating the patient to be sick.

TABLE 11. Comparison of ensembles schemes for Switzerland dataset.

Srno	Techniques	Accuracy	Sensitivity	Specificity	Fmeasure	Classification error	Kappa	Absolute Error	Relative Error	Correlation
1	Ensemble 1 (SVM+NB + Auto MLP)	89.39	53.87	55.06	54.46	10.61	0.087	0.171	17.10	0.052
2	Ensemble 2 (SVM+NB + NN)	91.21	60.59	63.86	62.18	8.79	0.240	0.165	16.49	0.152
3	Ensemble 3 (SVM+Perceptron+NN)	94.85	47.34	50.00	48.63	5.15	0.000	0.070	6.97	0.000
4	Ensemble 4 (DT+NB+ AutoMLP)	87.58	52.57	54.13	53.33	12.42	0.063	0.168	18.61	0.052
5	Ensemble 5 (SVM+NN + AutoMLP)	93.03	47.29	49.06	48.15	6.97	-0.027	0.076	7.55	0.000

TABLE 12. Comparison of ensembles schemes for Long Beach dataset.

Sr. no	Techniques	Accuracy	Sensitivity	Specifi-city	Fmeasure	Classification error	Kappa	Absolute Error	Relative Error	Correlation
1	Ensemble 1 (SVM+NB+Auto MLP)	74.21	63.37	55.99	59.45	25.79	0.139	0.277	27.72	0.171
2	Ensemble 2 (SVM+NB+NN)	72.11	57.23	53.24	55.17	27.89	0.082	0.298	29.82	0.144
3	Ensemble 3 (SVM+Perceptron +NN)	68.42	45.53	48.09	46.77	31.58	-0.010	0.323	32.28	0.077
4	Ensemble 4 (DT+NB+ AutoMLP)	75.00	64.91	56.35	60.32	25.26	0.155	0.268	26.84	0.195
5	Ensemble 5 (SVM+NN+ AutoMLP)	70.53	55.11	52.84	51.01	29.47	0.112	0.305	30.53	0.189

TABLE 13. Average comparison of ensemble schemes in terms of accuracy, sensitivity, specificity and f-measure.

Sr. no	Techniques	Accuracy (%)	Sensitivity (%)	Specificity (%)	F-Measure(%)
1	Ensemble 1 (SVM+NB+Auto MLP)	83.00	71.08	68.77	69.86
2	Ensemble 2 (SVM+NB+NN)	81.92	70.39	69.38	69.85
3	Ensemble 3 (SVM+Perceptron+NN)	81.67	62.52	64.02	63.25
4	Ensemble 4 (DT+NB+AutoMLP)	81.01	69.21	67.37	68.20
5	Ensemble 5 (SVM+NN+AutoMLP)	81.90	66.44	65.46	65.19

D. LONG BEACH DATASET

Long Beach data has 14 attributes and 200 instances. The missing values of rows have been replaced and the outliers have been detected and filtered out. This dataset has assigned 4 values (0, 1, 2, 3, and 4) to the goal attribute named ''num'' where 0 means absence of disease and implies that the patient is healthy and other values denote the presence of disease indicating the patient to be sick.

In the proposed technique, we have evaluated the performance of individual classifiers and ensemble vote schemes on 4 different heart disease data sets by using the following measures:

1) SENSITIVITY

Percentage of positive tuples that are correctly predicted by classifier as positive is called sensitivity [10]. It is calculated

TABLE 14. State of the art comparison of ensembles for heart disease datasets.

Author	Data sources and Instances used	Year	Techniques	Specificity	Sensitivity	Accuracy
Srinivas et al. [18]	Dataset in Attribute Relation File Format (ARFF) used from UCI repository (Not mentioned)	2010	NB	--	--	83.96%
Ghumbre et al. [21]	Medical records selected from automated diagnosis (214 instances)	2011	SVM RBF	88.50% 82.10%	84.06% 82.40%	85.05% 82.24%
Chen et al. [16]	1 dataset used from UCI repository (Cleveland dataset 303 instances)	2011	ANN	70%	85%	80%
Peter and Somasundurm. [15]	Dataset in attribute relation file format (ARFF) (675 instances)	2012	NB KNN DT NN	--	--	82% 74% 69% 78%
Jabbar et al. [17]	2 medical datasets used from UCI repository (Not mentioned)	2012	Genetic Algorithm	--	--	88%
Chitra and Sinivasagam. [19]	1 dataset used from UCI repository (270 patients)	2013	CNN SVM	87% 77.5%	83% 85.5%	85% 82%
Bashir et al. [24]	1 dataset used from UCI repository (Heart disease dataset 303 instances)	2014	Ensemble Vote approach	92.86%	73.68%	81.82%
Ismaeel et al. [26]	1 dataset used from UCI repository (Heart disease dataset 300 instances)	2015	Extreme Machine Learning (EML)	--	--	80%
Bashir et al. [11]	4 dataset used from UCI repository (1107 instances) 1 downloaded from ricco (209 instances)	2015	Ensemble Bagging Approach	82.87%	61.23%	81.87%
Verma et al. [25]	Clinical data collected from Indira Gandhi Medical Collage (IGMC) (335 instances)	2016	FURI MLP MLR C4.5	--	--	87.05% 90.28% 91.36% 85.6%
Khateeb et al. [20]	1 dataset used from UCI repository (Cleveland dataset 303 instances)	2017	K Nearest Neighbour	--	--	80%
Proposed Framework	4 dataset used from UCI repository 1. Cleveland dataset (303 instances)	2019	Ensemble 1 (SVM+NB+Auto MLP)	84.00%	84.80%	83.00%
	2. Hungarian dataset (294 instances)		Ensemble 2 (SVM+NB+Auto MLP)	82.30%	80.83%	83.49%
	3. Switzerland dataset (123 instances)		Ensemble 3 (SVM+Perceptron+NN)	47.34%	50.00%	94.85%
	4. Long Beach dataset (200 instances)		Ensemble 4 (DT+NB+Auto MLP)	64.91%	56.35%	75.00%

by the given formula:

$$sensitivity = \frac{True\ positives}{True\ positives + False\ positives} \tag{11}$$

2) SPECIFICITY

Relevant instance that are retrieved by a classifier is known as specificity [11]. It is calculated by given formula:

$$specificity = \frac{True\ negatives}{True\ negatives + False\ negatives} \tag{12}$$

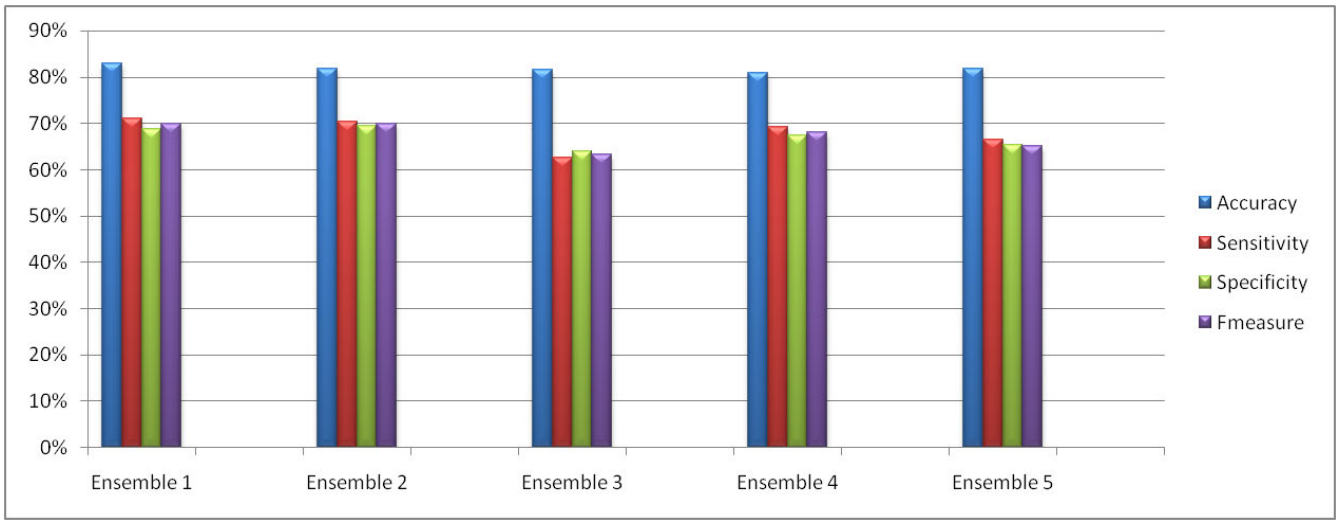


FIGURE 2. Average accuracy, sensitivity, specificity and f-measure comparison of five proposed ensemble vote schemes.

TABLE 15. ANOVA statistics for Cleveland heart disease dataset.

Ensembles	Source of Variation	Degree of Freedom DF	Sum of Squares SS	Mean Square MS	F-Stat	P-value
Ensemble 2	Between Groups	1	0.422442	0.422442	2.63	0.0065
	Within Groups	302	97.0165	0.160623		
Ensemble 3	Between Groups	1	1.485148515	1.485148515	9.7176	0.0019
	Within Groups	302	92.31023102	0.152831508		
Ensemble 4	Between Groups	1	2.259075908	2.259075908	14.247	0.0002
	Within Groups	302	95.77557756	0.158568837		
Ensemble 5	Between Groups	1	5.934675	5.934675	12.345	0.0013
	Within Groups	302	98.657864	0.1456757		

3) ACCURACY

True predictions made by proposed model on the test dataset is termed as accuracy [12]. It is calculated by given formula (13), as shown at the bottom of the page.

4) F-MEASURE

It is the weighted average of sensitivity and specificity [1]. It is calculated by using the following formula:

$$F - Measure = \frac{Sensitivity * Specificity}{Sensitivity + Specificity} \tag{14}$$

5) KAPPA STATISTICS

Kappa statistic is used to measure agreement for the predication between two raters. It is considered to be more accurate measure as compared to simple calculations as it uses the value of k that represents the possibility of agreement [34]. The formula for calculating k is given below:

$$k = \frac{p_0 - p_e}{1 - p_e} = 1 - \frac{1 - p_0}{1 - p_e} \tag{15}$$

where, p_0 represents observed agreement and p_e represents chance agreement probability. If there is a complete

$$Accuracy = \frac{True\ positives + True\ negatives}{True\ positives + True\ negatives + False\ positives + False\ negatives} \tag{13}$$

TABLE 16. ANOVA statistics for Hungarian dataset.

Ensembles	Source of Variation	Degree of Freedom DF	Sum of Squares SS	Mean Square MS	F-Stat	P-value
Ensemble 2	Between Groups	1	1.234321	1.234321	13.896	0.0007
	Within Groups	293	55.987657	0.124532		
Ensemble 3	Between Groups	1	1.365784	1.365784	11.876	0.0002
	Within Groups	293	57.76543	0.127651		
Ensemble 4	Between Groups	1	1.564324	1.564324	12.675	0.0003
	Within Groups	293	54.32456	0.123421		
Ensemble 5	Between Groups	1	1.232342	1.432344	14.213	0.0005
	Within Groups	293	56.32143	0.123212		

TABLE 17. ANOVA statistics for Switzerland dataset.

Ensembles	Source of Variation	Degree of Freedom DF	Sum of Squares SS	Mean Square MS	F-Stat	P-value
Ensemble 2	Between Groups	1	1.687542	1.687542	12.7865	0.00012
	Within Groups	122	72.56435			
Ensemble 3	Between Groups	1	1.54787	1.54787	11.7654	0.0001
	Within Groups	122	74.56752			
Ensemble 4	Between Groups	1	1.321232	1.321232	9.9876	0.00017
	Within Groups	122	71.78652			
Ensemble 5	Between Groups	1	1.45321	1.45321	12.6546	0.00011
	Within Groups	122	74.78976			

agreement between raters then k = 1 and if there is no agreement then k = 0, It is also possible that the value of k is negative which means no effective agreement between the raters.

6) CLASSIFICATION ERROR

Classification error is the percentage of incorrect predictions and it depends upon the number of samples classified incorrectly (false positive and false negative) [35]. It is calculated by the formula given below:

$$E_t = \frac{f}{n} * 100 \tag{16}$$

where E_t represents a single program, t depends on the no of samples f and n denotes the no of samples.

7) ABSOLUTE ERROR

Absolute error is the average absolute deviation from the measured value (predicted value) and actual value [36]. It is calculated by the formula:

$$\Delta x = x_0 - x \tag{17}$$

where Δx denotes absolute error, x_0 denotes the measured value and x denotes the actual value.

8) RELATIVE ERROR

It is the average absolute deviation from the measured value (prediction) and the actual value is divided by the actual value [37]. It is calculated by the formula given below:

$$\text{Re } lative\ error = \frac{Absolute\ error}{Actual\ value} \tag{18}$$

TABLE 18. ANOVA statistics for Long Beach dataset.

Ensembles	Source of Variation	Degree of Freedom DF	Sum of Squares SS	Mean Square MS	F-Stat	P-value
Ensemble 2	Between Groups	1	1.49	1.49	13.87	0.0001
	Within Groups	199	73.87654			
Ensemble 3	Between Groups	1	3.98765	3.98765	11.97	0.0009
	Within Groups	199	72.89653			
Ensemble 4	Between Groups	1	2.98762	2.98762	11.64	0.0005
	Within Groups	199	74.56251			
Ensemble 5	Between Groups	1	1.34	1.34	17.98	0.0002
	Within Groups	199	88.65432			

TABLE 19. Wilcoxon signed rank test for Cleveland heart disease dataset.

Cleveland heart disease dataset							
Ensemble 1	**Ensemble 5**	**Diff**	**Rank**	**Ensemble 1**	**Ensemble 2**	**Diff**	**Rank**
83.98	81.12	2.86	7	83.98	83.01	0.97	1
83.32	81.2	2.12	2	83.32	82.32	1	3
84.23	81.23	3	9	84.23	82.22	2.01	6
84.51	81.11	3.4	10	84.51	81.28	3.23	10
83.8	81.1	2.7	5	83.8	82.81	0.99	2
84.01	81.25	2.76	6	84.01	82.34	1.67	4
84.2	82.87	1.33	1	84.2	82.18	2.02	7
84.4	81.11	3.29	8	84.4	82.19	2.21	8
83.7	81.23	2.47	4	83.7	81.9	1.8	5
84.04	81.91	2.13	3	84.04	81.55	2.49	9
Ensemble 1	**Ensemble 3**	**Diff**	**Rank**	**Ensemble 1**	**Ensemble 4**	**Diff**	**Rank**
83.98	81.12	2.86	3	83.98	83.27	0.71	3
83.32	80.98	2.34	2	83.32	82.2	1.12	7
84.23	81.32	2.91	3	84.23	83.7	0.53	2
84.51	80.09	4.42	10	84.51	82.91	1.6	9
83.8	80.78	3.02	5	83.8	82.89	0.91	5
84.01	80.56	3.45	6	84.01	83.12	0.89	4
84.2	81.99	2.21	1	84.2	83.31	0.89	4
84.4	80.23	4.17	9	84.4	83.42	0.98	6
83.7	80.12	3.58	7	83.7	82.4	1.3	8
84.04	80.23	3.81	8	84.04	83.9	0.14	1

9) CORRELATION

Correlation is used to check the degree of relation between variables (quantitative or categorical variables). It uses a correlation coefficient for the prediction and to label attributes. The most commonly used correlation coefficient is Pearson Correlation Coefficient. If there is negative

TABLE 20. Wilcoxon signed rank test for Hungarian dataset.

Hungarian heart disease dataset							
Ensemble 1	Ensemble 5	Diff	Rank	Ensemble 1	Ensemble 2	Diff	Rank
81.42	80.01	1.41	5	81.42	79.92	1.5	3
81.59	80.3	1.29	4	81.59	79.76	1.83	9
81.76	80.07	1.69	8	81.76	79.54	2.22	10
81.73	81.03	0.7	2	81.73	79.98	1.75	8
81.54	80.05	1.49	6	81.54	79.91	1.63	4
81.61	80.12	1.49	6	81.61	79.93	1.68	5
81.32	80.1	1.22	3	81.32	79.86	1.46	2
81.05	80.52	0.53	1	81.05	79.79	1.26	1
81.64	80.11	1.53	7	81.64	79.92	1.72	6
81.95	80.04	1.91	9	81.95	80.21	1.74	7
Ensemble 1	Ensemble 3	Diff	Rank	Ensemble 1	Ensemble 4	Diff	Rank
81.42	76.82	4.6	3	81.42	76.78	4.64	1
81.59	76.64	4.95	8	81.59	76.93	4.66	2
81.76	76.98	4.78	6	81.76	76.45	5.31	4
81.73	77.75	3.98	1	81.73	76.23	5.5	5
81.54	76.86	4.68	4	81.54	75.08	6.46	10
81.61	77.03	4.58	2	81.61	76.02	5.59	7
81.32	76.51	4.81	7	81.32	75.09	6.23	9
81.05	76.09	4.96	10	81.05	76.07	4.98	3
81.64	76.91	4.73	5	81.64	76.09	5.55	6
81.95	77.01	4.94	9	81.95	75.97	5.98	8

correlation, positive correlation and neutral then value ranges will be from −1 to 1, 1 and 0 respectively [38]. It is calculated by the following formula:

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{N[\sum x^2 - (\sum x)^2 N \sum y^2 - (\sum y)^2]}} \tag{19}$$

where N represents total no of values, x represent values of first set of data and y represents values of second set of data.

V. RESULTS AND DISCUSSION

Experimentation is performed by using four benchmark heart disease datasets. These datasets are freely available at UCI data repository. 10 fold cross validation is applied on each dataset to divide them into training set and test set. Confusion matrix is then used to record the results of classifiers [11].

A large amount of patient's vitals can also be obtained from wearable devices which help in data collection and analysis [39]. These vital signs can be combined with electronic medical records (EMR) to improve the feature set [40]. UCI data repositories are also another source to provide benchmark heart disease datasets [41], [42].

A traditional confusion matrix is shown in Table 3.

Starting with the proposed approach, individual classifiers are evaluated on all dataset. In Table 4, we present the performance of individual classifiers for Cleveland data set. The results show that Support Vector Machine, Naive Bayes and AutoMLP have better accuracy, sensitivity, specificity and F-measure as compared to other classifiers. The top-3 classifiers also have low classification error, absolute error and relative error. This helps us to construct Ensemble 1 (SVM, NB and AutoMLP). In Table 5, we have evaluated the performances of different individual classifiers for Hungarian dataset. The results show that Support Vector Machine, Naive Bayes and Neural Network have better accuracy, sensitivity, specificity and F-measure as compared to other techniques. Moreover, SVM, NB and NN also have low classification error, absolute error and relative error. This results in the construction of Ensemble 2 (NN, NB and SVM).

In Table 6, we have evaluated the performance of different individual classifiers for Switzerland data set. The results show that Support Vector Machine, Perceptron and Neural Network have better accuracy as compared to

TABLE 21. Wilcoxon signed rank test for Switzerland dataset.

Switzerland heart disease dataset							
Ensemble 1	**Ensemble 5**	**Diff**	**Rank**	**Ensemble 1**	**Ensemble 2**	**Diff**	**Rank**
54.89	48.04	6.85	10	54.89	62.01	-7.12	10
54.75	48.2	6.55	8	54.75	62.09	-7.34	9
54.14	48.21	5.93	1	54.14	62.08	-7.94	2
54.35	48.13	6.22	5	54.35	62.09	-7.74	7
54.28	48.25	6.03	3	54.28	62.07	-7.79	5
54.26	48.24	6.02	2	54.26	62.18	-7.92	3
54.15	48.1	6.05	4	54.15	62.55	-8.4	1
54.72	48.12	6.6	9	54.72	62.08	-7.36	8
54.65	48.14	6.51	7	54.65	62.42	-7.77	6
54.41	48.07	6.34	6	54.41	62.24	-7.83	4
Ensemble 1	**Ensemble 3**	**Diff**	**Rank**	**Ensemble 1**	**Ensemble 4**	**Diff**	**Rank**
54.89	48.98	5.91	5	54.89	53.02	1.87	10
54.75	48.75	6	6	54.75	53.19	1.56	9
54.14	48.94	5.2	1	54.14	53.26	0.88	3
54.35	48.31	6.04	8	54.35	53.52	0.83	2
54.28	48.81	5.47	3	54.28	53.32	0.96	4
54.26	48.85	5.41	2	54.26	53.71	0.55	1
54.15	48.42	5.73	4	54.15	53.15	1	5
54.72	48.39	6.33	10	54.72	53.52	1.2	6
54.65	48.62	6.03	7	54.65	53.41	1.24	8
54.41	48.23	6.18	9	54.41	53.2	1.21	7

other classifiers. Moreover Neural Network also has better sensitivity, specificity and F-measure. Additionally, SVM, Perceptron and Neural Network also have low classification error. Furthermore, Perceptron and Neural Network have low absolute and relative error as compared to other classifiers. This concludes in the creation of Ensemble 3 (NN, SVM, Perceptron). Similarly, in Table 7, we have evaluated the performances of different individual classifiers for Long Beach data set. The results show that Decision Tree, Naïve Bayes and AutoMLP have better accuracy as compared to other techniques. Moreover, NB also has better sensitivity, specificity and accuracy. Additionally, DT, NB and AutoMLP also have low classification error. Furthermore, NB has low absolute and relative error as compared to other techniques. This results in the development of Ensemble 4 (DT, NB, AutoMLP).

Table 8 presents accuracy, sensitivity, specificity and F-measure of individual classifiers averaged over the four heart disease benchmark datasets. The results show that SVM, NN and Auto MLP have better accuracy as compared to other classifiers. This results in another combination of classifiers for our Ensemble 5 (NN, SVM, AutoMLP).

We have selected top-3 individual classifiers from each dataset on the basis of their high accuracy and combined them into ensembles. Hence five ensembles are generated as a result.

The ensembles are based on majority voting where the final class is selected which has highest number of votes from individual classifiers. It is also important to note that all results are computed using standard 10-fold cross validation.

Now, each of the ensembles, i.e. Ensemble 1-5, is evaluated on every benchmark dataset and the computed results for Ensemble 1, Ensemble 2, Ensemble 3, Ensemble 4 and Ensemble 5 for each heart disease data set are presented in Tables 9-12, respectively.

Finally, we have presented the average results of five ensemble vote schemes in terms of accuracy, sensitivity, specificity and F-Measure for four heart disease data set in in Table 13. The results show that Ensemble 1 has better accuracy for Cleveland and Hugarian dataset, Ensemble 3 has better accuracy for Switzerland data set whereas Ensemble 4 shows better results for Long Beach data set as compared to other techniques. We have combined the average results of five ensemble schemes in terms of accuracy, sensitivity,

TABLE 22. Wilcoxon signed rank test for Long Beach dataset.

Long Beach heart disease dataset							
Ensemble 1	Ensemble 5	Diff	Rank	Ensemble 1	Ensemble 2	Diff	Rank
59.42	51.76	7.66	2	59.42	54.99	4.43	6
59.46	50.93	8.53	4	59.46	55.28	4.18	2
59.87	50.97	8.9	7	59.87	55.11	4.76	10
59.84	50.98	8.86	6	59.84	55.24	4.6	8
59.41	51.29	8.12	3	59.41	55.02	4.39	5
58.32	51.37	6.95	1	58.32	55.61	2.71	1
59.39	50.32	9.07	8	59.39	55.07	4.32	4
59.74	50.09	9.65	9	59.74	55.13	4.61	9
59.49	51.37	8.12	3	59.49	55.29	4.2	3
59.59	51.02	8.57	5	59.59	55.01	4.58	7
Ensemble 1	Ensemble 3	Diff	Rank	Ensemble 1	Ensemble 4	Diff	Rank
59.42	46.62	12.8	6	59.42	59.39	0.03	7
59.46	46.98	12.48	2	59.46	59.44	0.02	6
59.87	46.76	13.11	8	59.87	61.73	-1.86	4
59.84	46.73	13.11	8	59.84	61.79	-1.95	3
59.41	46.71	12.7	5	59.41	59.4	0.01	5
58.32	46.45	11.87	1	58.32	60.93	-2.61	1
59.39	46.81	12.58	3	59.39	59.31	0.08	8
59.74	46.85	12.89	7	59.74	59.71	0.03	7
59.49	46.91	12.58	3	59.49	61.98	-2.49	2
59.59	46.93	12.66	4	59.59	59.57	0.02	5

specificity and f-measure. The results show that Ensemble 1 (SVM, NB, AutoMLP) has better average accuracy as compared to other ensembles schemes so this proposed ensemble can be used for heart disease diagnosis with high accuracy at real time. Graphical comparison of different ensembles is shown in figure 2. Each ensemble is evaluated on all four heart disease datasets and performance metrics are shown. The analysis of ensembles indicates that ensemble 1 has better results are compared to others. It has achieved an average accuracy of 83%.

Table 14 presents a state of the art comparison of proposed framework with other techniques in terms of specificity, sensitivity and accuracy for four benchmark heart disease dataset obtained from UCI repository [27]. We have compared them on the basis of performance and instances used by different authors.

A. STATISTICAL TESTS FOR COMPARING CLASSIFIER

The proposed Ensemble 1 is compared with other ensembles i.e Ensemble 2, Ensemble 3, Ensemble 4 and Ensemble 5 using two other statistical methods given as follows:

1) ANOVA STATISTICS

ANOVA (Analysis of Variance) Statistics is used to perform significance testing of ensembles [11]. The proposed ensemble i.e ensemble 1 is compared with all the other ensembles and p value is calculated. The results of ANOVA statistics are shown in tables 15-18. The p value and f-stat in datasets indicate that the results are statistically significant at 95% confidence interval. Therefore, the proposed ensemble (ensemble1) shows significant performance when compared with other ensemble classifiers.

2) WILCOXON'S SIGNED RANK TEST

Wilcoxon's signed rank test is a non-parametric method used to perform statistical comparison between classifiers [43]. Ensemble 1 is compared with other ensembles by using each fold as a trial. The difference in performance between pair of classifiers is compared for each fold. Critical value alpha =.05 is used to compare the value i.e minimum of sum of positive and negative ranks. If this minimum value is lower than alpha, indicates the rejection of null hypothesis which states that performance of two classifiers is same.

Tables 19-22 show Wilcoxon's signed rank test for four heart disease datasets. The results indicates that Ensemble 1 is significantly performs better as compared to other ensembles for most of the datasets and classifiers.

VI. CONCLUSION AND FUTURE WORK

In medical field, heart disease is one of the major diseases which can results in human death if it is not detected at early stages. The main objective of the proposed research is to predict heart disease more accurately independent of the underlying data. The proposed research focused on machine learning classifiers. Ensemble schemes were first proposed around 18 years ago in the field of data mining. This research paper proposed a majority voting based ensemble vote scheme for the accurate prediction of heart disease. A comprehensive empirical evaluation is conducted for possible combinations of individual classifiers in ensemble voting schemes. The five ensembles vote schemes (Ensemble 1-5) were tested on four benchmark heart disease dataset from UCI repository. The ensemble scheme performances were compared with each other and with other individual classifiers such as Neural Network, Decision Tree, Naïve Bayes, Support Vector Machine, Perceptron and AutoMLP. The average accuracy observed by the proposed ensemble vote scheme (Ensemble 1) was much better as compared to other techniques. It has achieved an average accuracy of 83%. Additionally, it has better sensitivity and f-measure as compared to other ensemble methods. The proposed framework can further be evaluated using more ensembles schemes like Adaboost, Bagging, Boosting and Stacking for heart disease prediction. Furthermore, this proposed framework can be analyzed for other disease prediction like Liver Disease, Breast Cancer and Diabetes Prediction.

REFERENCES

[1] S. Bashir, U. Qamar, and F. H. Khan, "IntelliHealth: A medical decision support application using a novel weighted multi-layer classifier ensemble framework," *J. Biomed. Inform.*, vol. 59, pp. 185–200, Feb. 2016.

[2] S. Bashir, U. Qamar, F. H. Khan, and M. Y. Javed, "MV5: A clinical decision support framework for heart disease prediction using majority vote based classifier ensemble," *Arabian J. Sci. Eng.*, vol. 39, no. 11, pp. 7771–7783, Nov. 2014.

[3] R. Das, I. Turkoglu, and A. Sengur, "Effective diagnosis of heart disease through neural networks ensembles," *Expert Syst. Appl.*, vol. 36, no. 4, pp. 7675–7680, May 2009.

[4] F. Zhu, P. Patumcharoenpol, C. Zhang, Y. Yang, J. Chan, A. Meechai, W. Vongsangnak, and B. Shen, "Biomedical text mining and its applications in cancer research," *J. Biomed. Informat.*, vol. 46, no. 2, pp. 200–211, Apr. 2013.

[5] T. Porter and B. Green, "Identifying diabetic patients: A data mining approach," in *Proc. AMCIS*, 2009, p. 500.

[6] K.-S. Leung, K. Hong Lee, J.-F. Wang, E. Y. T. Ng, H. L. Y. Chan, S. K. W. Tsui, T. S. K. Mok, P. C.-H. Tse, and J. J.-Y. Sung, "Data mining on DNA sequences of hepatitis b virus," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 8, no. 2, pp. 428–440, Mar. 2011.

[7] Mediacentre. (May 2017) *From World Health Organization*. Accessed: Apr. 4, 2018. [Online]. Available: http://www.who.int/mediacentre/factsheets/fs317/en/

[8] A. K. Sen, S. B. Patel, and D. P. Shukla, "A data mining technique for prediction of coronary heart disease using neuro-fuzzy integrated approach two level," *Int. J. Eng. Comput. Sci.*, vol. 2, no. 9, pp. 1663–1671, 2013.

[9] (Jun. 7, 2021). *Cardio Vascular Diseases (CVDs) (N.D.)*. From World Heath Organization. [Online]. Available: http://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[10] (2014). *Worldlifeexpectancy*. From World Health Ranikngs Live Longer Live Better. Accessed: Apr. 4, 2018. [Online]. Available: http://www.worldlifeexpectancy.com/cause-of-death/coronary-heart-disease/by-country/

[11] S. Bashir, U. Qamar, and F. H. Khan, "BagMOOV: A novel ensemble for heart disease prediction bootstrap aggregation with multi-objective optimized voting," *Australas. Phys. Eng. Sci. Med.*, vol. 38, no. 2, pp. 305–323, Jun. 2015.

[12] J.-K. Kim, J.-S. Lee, D.-K. Park, Y.-S. Lim, Y.-H. Lee, and E.-Y. Jung, "Adaptive mining prediction model for content recommendation to coronary heart disease patients," *Cluster Comput.*, vol. 17, no. 3, pp. 881–891, Sep. 2014.

[13] M. M. Abbasi and S. Kashiyarndi, *Clinical Decision Support Systems: A Discussion on Different Methodologies Used in Health Care*. Västerås, Sweden: Marlaedalen Univ. Sweden, 2006.

[14] S. A. Pattekari and A. Parveen, "Prediction system for heart disease using Naïve Bayes," *Int. J. Adv. Comput. Math. Sci.*, vol. 3, no. 3, pp. 290–294, 2012.

[15] T. J. Peter and K. Somasundaram, "An empirical study on prediction of heart disease using classification data mining techniques," in *Proc. Int. Conf. Adv. Eng., Sci. Manage. (ICAESM)*, Mar. 2012, pp. 514–518.

[16] A. H. Chen, S. Y. Huang, P. S. Hong, C. H. Cheng, and E. J. Lin, "HDPS: Heart disease prediction system," in *Proc. Comput. Cardiol.*, Sep. 2011, pp. 557–560.

[17] M. A. Jabbar, B. L. Deekshatulu, and P. Chandra, "Heart disease prediction system using associative classification and genetic algorithm," 2013, *arXiv:1303.5919*. [Online]. Available: http://arxiv.org/abs/1303.5919

[18] K. Srinivas, B. K. Rani, and A. Govrdhan, "Applications of data mining techniques in healthcare and prediction of heart attacks," *Int. J. Comput. Sci. Eng.*, vol. 2, no. 2, pp. 250–255, 2010.

[19] R. Chitra and V. Seenivasagam, "Heart disease prediction system using supervised learning classifier," *Bonfring Int. J. Softw. Eng. Soft Comput.*, vol. 3, no. 1, pp. 1–7, 2013.

[20] N. Khateeb and M. Usman, "Efficient heart disease prediction system using K-nearest neighbor classification technique," in *Proc. Int. Conf. Big Data Internet Thing (BDIOT)*, Dec. 2017, pp. 21–26.

[21] S. Ghumbre, C. Patil, and A. Ghatol, "Heart disease diagnosis using support vector machine," in *Proc. Int. Conf. Comput. Sci. Inf. Technol. (ICCSIT)*, Dec. 2011, pp. 84–88.

[22] T. Helmy, S. M. Rahman, M. I. Hossain, and A. Abdelraheem, "Non-linear heterogeneous ensemble model for permeability prediction of oil reservoirs," *Arabian J. Sci. Eng.*, vol. 38, no. 6, pp. 1379–1395, Jun. 2013.

[23] J. Maroco, D. Silva, A. Rodrigues, M. Guerreiro, I. Santana, and A. de Mendonça, "Data mining methods in the prediction of dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests," *BMC Res. Notes*, vol. 4, no. 1, p. 299, Dec. 2011.

[24] S. Bashir, U. Qamar, and M. Younus Javed, "An ensemble based decision support framework for intelligent heart disease diagnosis," in *Proc. Int. Conf. Inf. Soc. (i-Soc.)*, Nov. 2014, pp. 259–264.

[25] L. Verma, S. Srivastava, and P. C. Negi, "A hybrid data mining model to predict coronary artery disease cases using non-invasive clinical data," *J. Med. Syst.*, vol. 40, no. 7, p. 178, Jul. 2016.

[26] S. Ismaeel, A. Miri, and D. Chourishi, "Using the extreme learning machine (ELM) technique for heart disease diagnosis," in *Proc. IEEE Canada Int. Humanitarian Technol. Conf. (IHTC)*, May 2015, pp. 1–3.

[27] D. Nikovski, "Constructing Bayesian networks for medical diagnosis from incomplete and partially correct statistics," *IEEE Trans. Knowl. Data Eng.*, vol. 12, no. 4, pp. 509–516, Jul. 2000.

[28] C. Nunes, H. Langet, M. De Craene, O. Camara, B. Bijnens, and A. Jonsson, "Decision tree learning for uncertain clinical measurements," *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 9, pp. 3199–3211, Sep. 2021.

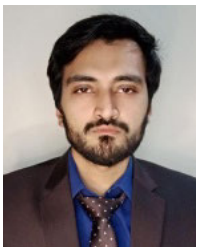
[29] S. Zhao, M. Jiang, B. Qin, T. Liu, C. Zhai, and F. Wang, "Structural and textual information fusion for symptom and disease representation learning," *IEEE Trans. Knowl. Data Eng.*, early access, Nov. 20, 2020, doi: 10.1109/TKDE.2020.3039469.

[30] J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan, and A. Saboor, "Heart disease identification method using machine learning classification in E-healthcare," *IEEE Access*, vol. 8, pp. 107562–107582, 2020.

[31] A. Shamsi, H. Asgharnezhad, S. S. Jokandan, A. Khosravi, P. M. Kebria, D. Nahavandi, S. Nahavandi, and D. Srinivasan, "An uncertainty-aware transfer learning-based framework for COVID-19 diagnosis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 4, pp. 1408–1417, Apr. 2021.
[32] (Jun. 7, 2018). *Heart Disease Data set. (N.D.)*. UCI. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Heart+Disease
[33] S. Mirjalili, "Evolutionary multi-layer perceptron," in *Evolutionary Algorithms and Neural Networks*. Cham, Switzerland: Springer, 2019, pp. 87–104.
[34] A. De Raadt, M. J. Warrens, R. J. Bosker, and H. A. L. Kiers, "Kappa coefficients for missing data," *Educ. Psychol. Meas.*, vol. 79, no. 3, pp. 558–576, Jun. 2019.
[35] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6402–6413.
[36] C. J. Willmott and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance," *Climate Res.*, vol. 30, no. 1, pp. 79–82, Dec. 2005.
[37] B. Thiam, "Relative error prediction in nonparametric deconvolution regression model," *Statistica Neerlandica*, vol. 73, no. 1, pp. 63–77, Feb. 2019.
[38] F. H. Khan, U. Qamar, and S. Bashir, "A semi-supervised approach to sentiment analysis using revised sentiment strength based on SentiWordNet," *Knowl. Inf. Syst.*, vol. 51, no. 3, pp. 851–872, 2017.
[39] F. Ali, S. El-Sappagh, S. M. R. Islam, A. Ali, M. Attique, M. Imran, and K.-S. Kwak, "An intelligent healthcare monitoring framework using wearable sensors and social networking data," *Future Gener. Comput. Syst.*, vol. 114, pp. 23–43, Jan. 2021.
[40] F. Ali, S. El-Sappagh, S. M. R. Islam, D. Kwak, A. Ali, M. Imran, and K.-S. Kwak, "A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion," *Inf. Fusion*, vol. 63, pp. 208–222, Nov. 2020.
[41] S. Mishra, A. Dash, P. Ranjan, and A. K. Jena, "Enhancing heart disorders prediction with attribute optimization," in *Advances in Electronics, Communication and Computing*. Singapore: Springer, 2020, pp. 139–145.
[42] S. Sahoo, M. Das, S. Mishra, and S. Suman, "A hybrid DTNB model for heart disorders prediction," in *Advances in Electronics, Communication and Computing*. Singapore: Springer, 2021, pp. 155–163.
[43] A. Bifet, G. de Francisci Morales, J. Read, G. Holmes, and B. Pfahringer, "Efficient online evaluation of big data stream classifiers," in *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2015, pp. 59–68.

SABA BASHIR received the Ph.D. degree in software engineering from the College of E&ME, National University of Science and Technology, Pakistan. She is currently an Assistant Professor with the Computer Science Department, Federal Urdu University of Arts, Science and Technology, Islamabad, Pakistan. She is a Certified Engineer. She has published many research articles in top-tier peer-reviewed international journals in the field of data mining/machine learning. Her research interests include predictive systems, health informatics, artificial intelligence, and machine learning.

ABDULWAHAB ALI ALMAZROI received the M.Sc. degree in computer science from the University of Science, Malaysia, and the Ph.D. degree in computer science from Flinders University, Australia. He is currently working as an Assistant Professor with the Department of Information Technology, College of Computing and Information Technology at Khulais, University of Jeddah, Saudi Arabia. His research interests include parallel computing, cloud computing, wireless communications, and data mining.



SUFYAN ASHFAQ received the B.S. degree (CS) from the Federal Urdu University of Arts, Science and Technology (FUUAST), Islamabad, Pakistan, and the B.Sc. degree from the University of Punjab, Lahore. He is currently pursuing the M.S. degree in computer science with FUUAST. His research interests include machine learning and deep learning.

ABDULALEEM ALI ALMAZROI received the Ph.D. degree in computer science in the field of computer networks from Universiti Teknologi Malaysia, in 2016. He is currently an Assistant Professor with the Applied College, Northern Border University, Saudi Arabia. His research interests include computer networks, cyber security, data mining, and artificial intelligence.

FARHAN HASSAN KHAN received the Ph.D. degree from the National University of Science and Technology (NUST), Pakistan. He is currently a Project Director of the National Research & Development Organization, Pakistan. He is also a Deputy Director (Research) of the Knowledge and Data Science Research Center, College of E&ME, NUST, where he is heading the NLP Group. He has more than 16 years of professional experience in software industry as well as academia. His research interests include data/text mining and health informatics supported by numerous publications in top-tier international journals and conferences. He is also serving as a member for editorial boards and a reviewer at various internationally recognized venues. He was also awarded the President's Gold Medal for excellent academic performance from NUST.